Prague University of Economics and Business
Faculty of Informatics and Statistics

Bayesian analysis of time-varying volatility models

MASTER THESIS

Study program: Statistics
Specialization: Econometrics

Author: Bc. et Bc. Michal Lauer
Supervisor: Ing. Miroslav Plašil, Ph.D.

Prague, June 2025

Acknowledgements

I would like to thank my supervisor for their guidance in the world of bayesian statistics.

Abstrakt

Abstrakt [CZE]

Klíčová slova

Bayesovská statistika, Finanční časové řady, ARCH, GARCH, SV

Abstract

Abstrakt [ENG]

Keywords

Bayesian statistics, Financial time series, ARCH, GARCH, SV

Table of contents

List of Figures

List of Tables

List of source codes

List of abbreviations

CDS Cross-Sectional Data
TSD Time Series Data
AR Autoregressive model
MA Moving-Averages model
ARIMA Autoregressive Integrated Moving-Averages model
ARCH Autoregressive Conditional Heteroscedasticity
MLE Maximum Likelihood Estimation
GARCH Generalized Autoregressive Conditional Heteroscedasticity
EGARCH Exponential generalized auto regressive conditional heteroscedasticity
GARCH-M GARCH-in-Mean
SV Stochastic Volatility
LLN Law of Large Numbers
PI Percentile Interval
HDI Highest Density Interval
MCMC Markov Chain Monte Carlo
ESS Effective Sample Size
MA Metropolis Algorithm
MHA Metropolis-Hastings Algorithm
HMC Hamiltonian Monte Carlo
NUTS No-U-Turn Sampler
VI Variational Inference
KLD Kullback–Leibler Divergence

Introduction

This thesis combines two domains which are growing in their popularity, complexity and importance. The roots of bayesian statistics go back to the late 18th a and early 19th century and are summarized in (Hebák, 2013). The core tool of bayesian statistics - the bayes' theorem - is attributed to mathematician and Thomas Bayes. His famous work *An Essay Towards Solving a Problem in the Doctrine of Chances*, which was release post-mortem by his friend Richard Price is the first work that mentions and defines the theorem. His work was extended by Pierre-Simon Laplace, who was able to capture and explain more meaningfully the nuances of a the choice of apriori distribution or the problem of estimating an unknown parameter of binomial distribution.

In the late 19th century and most of the 20th century, frequentist statistic became popular and subjective, bayesian interpretation of probability has been neglected. During this period, the famous statistician Sir Ronald Alymer Fisher contributed to frequentist statistics with his work on parameters and their properties, such as unbiasedness or consistency. His most influential work was in statistical inference and the development of tools to conduct hypothesis testing. Other than a great statistician and mathematician, Fisher was also an advocate of frequentist and objective approach.

Fisher famously criticised bayesian statistics, then referred to as *inverse* statistics, as a *' inverse probability, which like an impenetrable jungle arrests progress towards precision of statistical concepts'*. Interestingly, his stance has not been always consistent and in his early work on Maximum likelihood estimation, he pleads to *' to having based his argument upon the principle of inverse probability'*. (Zabell, 2022) further explains that Fisher's issue was primarily with universally uniform priors, which are not scale invariant. However, universally uniform priors are not standard in modern bayesian statistics and popular textbooks do not encourage this approach. (Kruschke, 2015, Chap. 10.6) talks about extreme sensitivity to prior knowledge about estimated parameters and in model comparison. He also suggests that many statisticians support using other uninformative priors which might be more appropriate for specific models. So while Fishers critique was legitimate, it may not hold in current bayesian settings.

With increased computational availability, bayes statistics started to become more popular. Introduction of sampling methods that are capable of generating samples from a posterior distribution that is not analytically known lead to models that can be more complicated while still usable in research areas. Currently, bayesian statistics is utilized in many fields. (Ashby, 2006) examines the state of bayesian statistics in medicine towards the end of 20th century. One of the conclusions is that bayes has *'now permeated all the major areas of medical statistics, including clinical trials, epidemiology, meta-analyses and evidence synthesis, spatial modelling, longitudinal modelling, survival modelling, molecular genetics and decision-making'*. (Barber, 2012) draws connection between bayesian reasoning and machine learning. The similarities between bayesian thinking and quantum analysis is described in

(Timpson, 2008).

Analysis of volatility of financial markets in the current digital world is important and (Liu, 2024) provides at least 2 domains where this might be beneficial. Assessing the volatility of markets can help with portfolio optimization and risk management. Studies found that the inclusion of GARCH models, which are the topic of this thesis, has lead to a 20 % decrease in portfolio volatility. This further helps management and traders mitigate potential losses and improve decision making. Another popular domain are derivative markets for futures or options where setting the right price is crucial. Historically, option pricing uses methods such as the Black Scholes model, which assumes that the variance of the underlying asset is constant. This assumption does not always hold in practice, and the need to support heteroskedasticity in financial time series arose.

The effect of financial volatility can overlap with other fields. ('Quantitative Analysis on Economic and Financial Factors behind International Students Tourism (Study Destination Choice): Evidence of China', 2019) investigated a relationship between the China's exchange rate against the US dollar and found that it may have some effect on international student tourism between the US and China. A larger study that explores the relationship between country credit rating, political risk, financial risk, economic risk and a composite risk rating conducted in (Erb et al., 1996). The conclusion suggest that there is some relationship between different risks, especially in financial and economic ratings.

The bridge between novel bayesian statistical thinking and financial markets that can affect everyday life is interesting in several ways. First, some advanced models that are capable of measuring volatility in financial markets are hard to estimate and bayesian methods naturally arise, especially in the case of stochastic volatility. Furthermore, bayesian statistics offer information not only point and interval estimates, but in a full posterior distribution from which arbitrary points, quantiles and intervals are computable. This supports advanced decision making and offers more information about the estimated quantities. The probabilistic interpretation that can also be more natural and helpful rather than 'imaginary repetitions in identical conditions'.

This thesis has two main goals. In the theoretical part, a connection between financial markets and bayesian models is drawn using frequentist methods for stochasticity modeling. Methods are introduced, derived and translated from frequentist world to the bayesian world. AThese bayesian models are then used in practical part on time series from financial markets, where the underlying volatility is captured. The structure is divided into three main chapters. First chapter explores financial time serie and their statistical characteristics. Classical models which are used throughout the thesis are also defined and described. The second part focuses on bayesian statistics. Initially, the ideas behind bayesian statistics are introduced as well as model definition, statistical inference and the creation of posterior distributions. After that, different sampling methods are used which are crucial in models where analytical posterior distribution is not derived. Finally, the connection between volatility models and bayesian statistics is drawn using available studies and surveys. The last chapter focuses on data

analysis in R (R Core Team, 2023) and Stan (Carpenter et al., 2017).

1. Analysis of time series data

Analysis of time series can be more complicated that analysis of cross sectional data because the time dimension that is inherently introduced. Methods on data without time dependence often assume random sampling and identical, independent distribution of every data point. An example can be the MLR3 requirement defined in (Wooldridge, 2020) for Multiple Linear Regression estimated with Ordinary Least Squares. It states that in order to get unbiased estimates $\hat{\theta}$ of the true population parameters θ, every data points needs to be independent. In time series data, a point at time t is dependent on values up-to time $t-1$ and this requirement is automatically violated.

(Hindls et al., 2018) states that analysis of time series data is the most important analysis for economic data. All economic data is *somehow* dependent on its past values and this dependence need to be reflected. One way of analyzing a time series is by forecasting future values. This can be done using either point estimates or confidence intervals. Another approach, that is considered in this thesis, is the forecast of future volatility. Because a lot of popular techniques, such as the Black–Scholes model, assume homoscedasticity, it is only natural to model the underlying volatility.

1.1 Time series

Assume a data sequence y that can be denoted using an index t. These indexes serve as ordering in time from the beginning $t = 1$ to end $t = T$. This data sequence denotes a time series

$$y_t \text{ for } t = 1, 2, \dots, T.$$

Because TSD are generally dependent, their complete description can be provided by a joint probability distribution function F. For indexes $t_1, t_2, \dots, t_n$ where $n \in \mathbb{Z}$ and constants $c_1, c_2, \dots, c_n$, a general joint distribution function (Shumway & Stoffer, 2017, Chap. 1.3) is defined as

$$F_{t_1,t_2,\dots,t_n}(c_1, c_2, \dots, c_n) = \Pr(x_1 \leq c_1, x_2 \leq c_2, \dots, x_n \leq c_n).$$

Because of the high dimensionality and complexity, these functions usually cannot be evaluated and the marginal distributions,

$$f_t(x) = \frac{\delta F_t(x)}{\delta x},$$

are often more informative. In addition to marginal distributions, a time series at time t can be described by the mean function. Provided it exists, the mean function at time t is defined by (Shumway & Stoffer, 2017, definition 1.1) as

$$\mu_{x_t} = E(x_t) = \int_{-\infty}^{\infty} x f_t(x) dx.$$

To measure the dependence of time series, autocovariance function offers a measure of linear dependence between two different points t and s. Such function is defined by (Shumway & Stoffer, 2017, definition 1.2) as

$$\gamma_x(s,t) = E\left[(x_s - \mu_s)(x_t - x_s)\right].$$

If $t = s$, autocovariance measure the variance of time series at time t. If the function is the same for all indexes t, that is $\gamma(t) = \gamma(s)$ $t, s = 1, \ldots, n$, the time series has a constant variance and is called homoskedastic. Otherwise, the series exhibits different variability at different times, and is called heteroskedastic.

1.1.1 Stationarity

In addition to the mentioned decomposition and characteristics, TSD can be described based on how the data distribution changes in time. If the probability distribution is identical at every time point t, the distribution is strictly stationary. This means that shifting in any direction by any distance $h \in \mathbb{Z}$ does not change the underlying distribution and

$$\Pr(x_t \leq c) = \Pr(x_{t+h} \leq c).$$

for $c \in \mathbb{R}$ (Shumway & Stoffer, 2017, definition 1.4). Same distribution function at any point t implies that

$$\mu_s = \mu_t,$$

if the mean function exists. The variance is also the same and the linear dependence, measured by covariance, depends only on the time shift h. If the function exists, this relationship can be written as

$$\gamma(t,s) = \gamma(t+h, s+h).$$

Strict stationarity is often too strong to be applied in practical examples. It also defines that the distribution is the same for *all* time points t, which is often impossible to assess from a

single data set. Instead of a strictly stationar process, it is often sufficient to have a weakly stationary process. (Shumway & Stoffer, 2017, Def. 1.7) defines a weakly stationary time series such that

1) the mean value μ does not depend on time t, and
2) the autocovariance function depends only on the time shift h.

The main difference between strict and weak stationarity is that strong stationarity assumes identical distribution functions while weak stationarity assumes only the first two moments to be identical. This also means that if a series is strictly stationar, it is also weakly stationar. But the opposite is not true. Because of the hard implications imposed by strict stationarity, this thesis will always refer to weak stationarity under the term 'stationarity'.

As an example, Figure 1.1 taken from (Shumway & Stoffer, 2017, Fig. 1.1) shows some of these properties on real data.

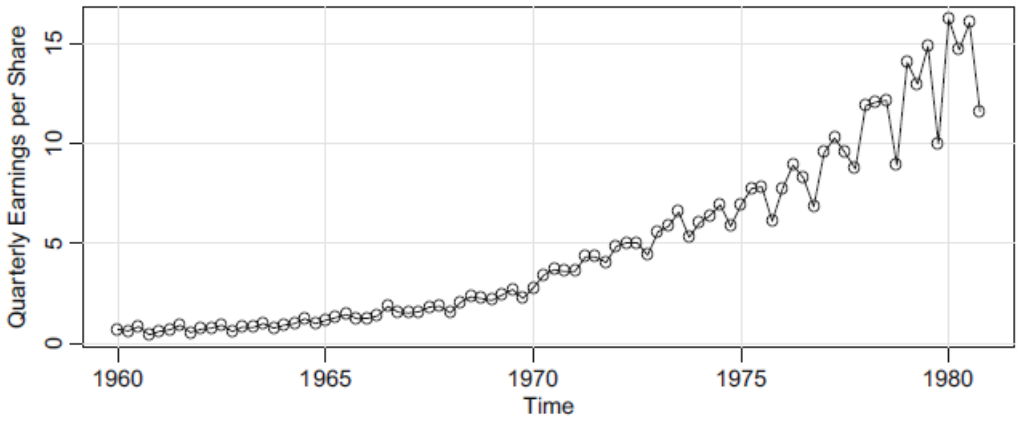

Figure 1.1: Johnson & Johnson quarterly earnings per share, 84 quarters, 1960-I to 1980-IV (Shumway & Stoffer, 2017)

At first glance, it can be observed that the series has a positive trend with cyclicality. As time goes, the data are more variable and thus, the series suffers from heteroskedasticity. Because the mean μ is most probably not time-invariant, the series can not be stationary under neither definition. Such visual analysis is often beneficial before any rigid statistical analysis (Shumway & Stoffer, 2017, Chap. 1).

Stationarity is one of the most important properties that in TSD analysis. As (Tsay, 2005, Chap. 2.1) puts is, *'The foundation of time series analysis is stationarity'*. A lot of models that are frequently used for time series data analysis require stationarity (or some transformation that is stationary). An example can be the autoregressive model (AR) which assumes that the observed value at time t is linearly dependent on it's past values. If the necessary condition for stationarity[1] is not met, the model should not be used because the time series can explode. Forecasting of such time series would be very difficult. An extension of such models is an ARIMA model, that combines the AR process with moving averages

[1]Generalized condition for $AR(p)$ process is defined in (Tsay, 2005, Chap. 2.4.1)

model (MA). (Tsay, 2005, Chap. 2.5). This model employs integration, where the data is transformed by differencing. A time series is said to be stationary if after d differences, the transformed series itself is stationary. Further properties of AR, MA and ARIMA as well as other, more complex and traditional models are described in (Tsay, 2005).

Another issue that is often met in TSD analysis is heteroskedasticity. Different variability at different times implies that the time series is not stationar and specific approach must be considered. Heteroskedasticity is also very often connected to financial data, because one can find periods of high volatility markets and low volatility markets (Tsay, 2005, Chap. 3). Models that can handle such time series are called Conditional Heteroskedastic Models and will be the main topic of discussion in this thesis.

1.2 Financial time series

Financial data and time series analysis can be loosely defined as *analysis that is concerned with the theory and practice of asset valuation over time* (Tsay, 2005, Chap. 1). Key feature that distinguishes financial TSD from other domains is volatility, which originates from both not unified theory and variable data. A simple term, such as asset volatility, can have different definitions, which complicates statistical analysis and empirical research. The volatility of data itself is also hardly measurable and data that changes quickly (such as returns or closing price) can experience different periods of different volatility (Tsay, 2005, Chap. 1).

In financial statistical analysis, it is often beneficial to work with returns and not pure prices. In fact, most of the current research in finance is focused on returns (Tsay, 2005, Chap. 1). They offer scale-free metric that measure how well a trader performs, no matter the volume. They have also good statistical properties which are beneficial in TSD analysis. One of those is stationarity that describes constant first and second moments (Campbell et al., 1998, Chap. 1.4).

There are different methods on how to calculate returns (Tsay, 2005, Chap. 1.1). The simplest returns at time t are called simple gross returns, and are denoted as

$$1 + R_t = \frac{P_t}{P_{t-1}}, \tag{1.1}$$

where P_t denotes the price of the underlying asset at time t and R_t the net return. Equation 1.1 can also be interpreted as a simple growth index between two consecutive time periods. Extension of this are log returns, which can be defined as

$$r_t = \ln(1 + R_t) = \ln\left(\frac{P_t}{P_{t-1}}\right) = \ln P_t - \ln P_{t-1} = p_t - p_{t-1}, \tag{1.2}$$

where p_t denotes the natural logarithm of price at time t. This transformation is beneficial

in two ways. First, it is simple to write the compound growth in time as a simple sum. Considering growth the past h records, it can be computed that

$$\ln\left(\frac{P_{t-h+1}}{P_{t-h}} * \cdots * \frac{P_{t-1}}{P_{t-2}} * \frac{P_t}{P_{t-1}}\right) = \ln\left(\frac{P_{t-h+1}}{P_{t-h}}\right) + \cdots + \ln\left(\frac{P_{t-1}}{P_{t-2}}\right) + \ln\left(\frac{P_t}{P_{t-1}}\right) = \sum_{i=t-h+1}^{t} p_i.$$

Another important property touches statistical analysis and the fact that natural logarithm is often easier to analyze (Tsay, 2005, Chap. 1.1).

(Cont, 2007, Chap. 1) provides an overview of some statistical properties that have been observed on financial returns. Volatility of a financial asset is positively correlated with it's traded volume. Volatility clustering is also usually observed and large changes are typically followed by large changes, while small changes are typically followed by small changes. These changes, however, can be in a different direction and relatively high returns or losses can be observed together. While these returns are typically not correlated, their absolute value experiences autocorrelation with slow decay. Finally, unconditional distribution of returns displays heavy tails.

If a financial time series experiences volatility clustering, it is inherently not stationary, as it's variance is not constant. As (Cont, 2007, Chap. 1) suggests, this issue has lead to a development of new models that are specifically designed to handle periods of high and low time series volatility.

1.3 Conditional heteroskedastic models

To combat heteroskedasticity and model volatility that is time-variant, models of conditionally homoskedastic models have been developed. One of the most popular and traditional models assumes that the volatility at time t can be deterministically described by its past values. Typically, these models are represented by autoregressive functions of squared residuals and historical variance. Although harder to estimate, this method has proven to be beneficial especially in TSD analysis (Fernández & Rodríguez, 2020).

A newer approach assumes that the volatility at time t is modelled through a latent variable that is also autoregressively updated. This method has become popular because of it's ability to capture more complex relationships than deterministic models. The estimation of such models is typically harder, because analysts need to estimate latent variables that are not observed. Some specialized methods have been developed to counter this issue, and the use of bayesian simulations is popular choice. Some authors argue that the additional stochasticity through latent variables improves model fitting and predictions (Fernández & Rodríguez, 2020), while others argue that the out-of-sample predictions do not differ much (Tsay, 2005, Chap. 3.12).

1.3.1 ARCH Model

Log returns in financial TSD analysis are serially uncorrelated, but dependent. Figure 1.2, taken from (Tsay, 2005, Fig. 3.1), shows log returns of Intel Corporation from January 1973 to December 2003.

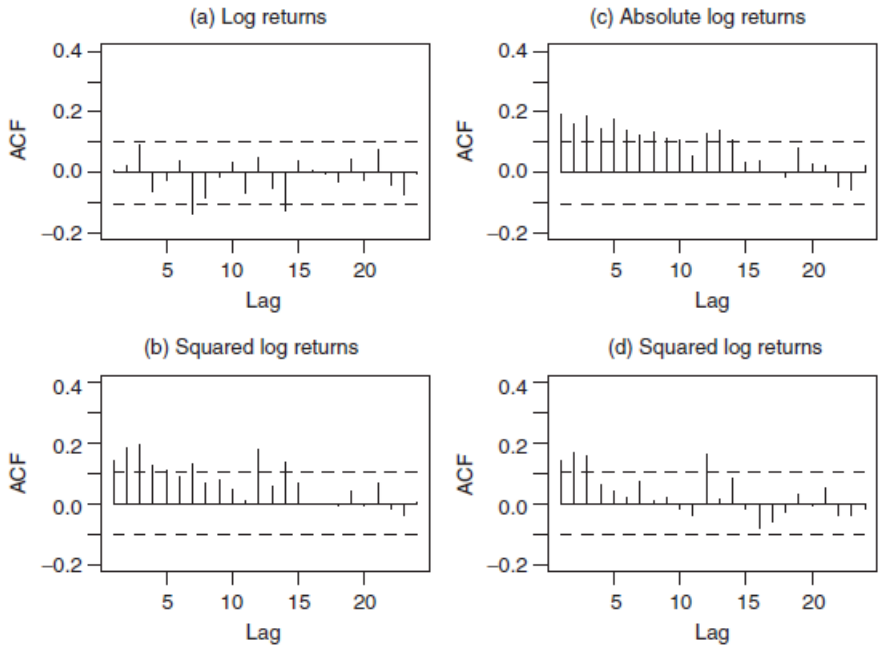

Figure 1.2: Various transformation of log returns of Intel Corporation (Tsay, 2005)

The raw log returns on the top left show that there is no serial autocorrelation, except for some small spikes that could be due to randomness. The other two transformations (absolute values, squares) exhibit some type of dependence. Since the volatility is not independent, it is viable to try to use the past values for statistical modeling.

The log returns could be described by

$$p_t = \mu_t + a_t,$$

where p_t are the log returns at time t, μ_t is so-called mean equation at time t and a_t is the shock[2] at time t. The mean equation is a model that describes the expected value at time t, given all information up to $t-1$. It can be a simple $AR(p)$ or $ARIMA(p,d,q)$ process that is stationary. Shock at time t describes the volatility at time t with σ_t and a random noise. This shock is the main focus of volatility modeling (Tsay, 2005, Fig. 3.2).

Let shock a_t be defined as

$$a_t = \sigma_t \epsilon_t,$$

[2]Some authors uses also the word *innovation*.

where ϵ_t represents a sequence of identically and independently distributed random variable with $\mu = 0$ and $\sigma^2 = 1$ and σ_t is an autoregressive process

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{m} \alpha_i a_{t-i}^2. \tag{1.3}$$

Put together, the model is called Autoregressive conditional heteroskedasticity (ARCH) model of order m. Equation 1.3 states that the variance at time t is affected by squared residuals at the previous m times. This implies that, if the volatility has been high, the volatility at time t is expected to be high as well. On the contrary, in periods of low volatility, the expected volatility at time t is rather small. It is important to realize that high volatility does not necessarily imply high realized variability. The probability of very high or very low returns is greater, but not guaranteed (Tsay, 2005, Chap. 3.4).

The random variable ϵ_t is often assumed to follow standard normal distribution. If the time series experiences heavy tails, the student distribution might be more suitable. Since the expected value of such distributions are zero, the expected return at time t is not affected and by (Tsay, 2005, Chap. 3.4.1), it holds that

$$E(p_t) = E(\mu_t) + E(a_t) = E(\mu_t).$$

Because σ^2 is variance, it must be greater than zero and such, $\alpha_0 > 0$ and $\alpha_i \geq 0$ for $i \in \mathbb{N}$.

In frequentist statistics, there are several ways how to estimate the correct order of ARCH(m). One of them is the Ljung-Box statistic $Q(m)$ computed from the squared residuals a_i^2. Another popular method is the Lagrange multiplier test, that is identical to the usual F-test in multiple regression. To estimate such model, maximum likelihood estimation (MLE) with different likelihood functions is commonly used. Because these thesis focuses primarily on bayesian methods and estimation, these methods will not be leveraged and further description and application is offered in (Tsay, 2005, Chap. 3).

To use the ARCH(m) model for forecasting, it is necessary to estimate historical residuals. One-step forecast for ARCH(m) model is made up of the mean equation that is, in general, not affected by the forecasted volatility. After computing historical predictions $\hat{\mu}_t$, it is necessary to compute residuals, which represent the historical shock as

$$\hat{a}_t = y_t - \hat{\mu}_t.$$

These residual shocks are then used in the ARCH(m) model to forecast future volatility as

$$\hat{\sigma}_{t+1}^2 = \hat{\alpha}_0 + \sum_{i=1}^{m} \hat{\alpha}_i \hat{a}_{t-i+1}^2.$$

Forecasting equation can be generalized for arbitrary ℓ-step ahead prediction (Tsay, 2005, Chap. 3.4.3)

(Tsay, 2005, Chap. 3.4.2) defines downsides of such models. One of them is the assumption that positive and negative spikes of residuals have the same effect on the current volatility. Research has found that this is not true and squared historical shocks might not correctly reflect the true shock. ARCH models also tend to overestimate volatility in cases where a single shock has been observed. Simple ARCH models respond slowly to such quick changes and might not offer the best performance.

1.3.2 GARCH Model

The variance for ARCH(m) model given in Equation 1.3 is simple and effective, but sometimes not adequate for real data. General autoregressive conditional heteroscedasticity (GARCH) model, proposed by Boreslav (Bollerslev, 1986), are more flexible and offer better long-term memory for forecasting (AL-Najjar, 2016). GARCH models extend ARCH shock by lagged estimated variance. In general, the GARCH model of order p and q can be written as

$$p_t = \mu_t + a_t,$$

where the shock element a_t is made up by deterministic component σ_t and stochastic component ϵ_t. The main difference is in σ_t, which is given by (AL-Najjar, 2016, Eq. 2) as

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{p} \alpha_i a_{t-i}^2 + \sum_{j=1}^{q} \beta_j \sigma_{t-j}^2. \tag{1.4}$$

The main difference between ARCH(p) model and GARCH(p, q) model is the sum of past conditional variances given in Equation 1.4. In addition to the ARCH constrains, for the GARCH β coefficients it must hold that $\beta_j \geq 0$ and $\sum_{i=1}^{\max(p,q)}(\alpha_i + \beta_i) < 1$. This condition implies that the shock a_t is finite and conditional variance σ^2 evolves over time (Tsay, 2005, Chap. 3.5).

New β coefficients introduce complexity and assessing the order of GARCH model is a complex task. In practice, only lower orders are used. To estimate a GARCH model, a recursive method is viable which first estimates a simple ARCH(p) model. From the estimated residuals a_t and conditional variance σ_t^2, new GARCH(p, q) model is computed. This creates new residuals a_t and conditional variance σ_t^2, which can be used to fit a new GARCH(p, q) model. This procedure is repeated until a stopping criterion is met. To forecast new values, the same procedure as for the ARCH model can be used and arbitrary ℓ-step ahead prediction is available (Tsay, 2005, Chap. 3.5)

Although the GARCH(p, q) model has some drawbacks, extensions have been developed to improve volatility forecasts. One popular extension is the Exponential generalized auto regressive conditional heteroscedasticity (EGARCH) model, that is able to capture asymmetric shocks. This means that positive and negative jumps in returns can have various effect on the implied forecasted volatility. Another extensions, GARCH-in-Mean (GARCH-M), uses the calculated conditional volatility from GARCH model as a predictor to forecast returns. Overview of other, more specific extensions that might be appropriate in specific domains, are covered in (AL-Najjar, 2016) or (Tsay, 2005).

1.3.3 Model of Stochastic volatility

Another approach to volatility modeling is to leave the assumption that volatility at time t can be deterministically described by previous shocks and conditional variance. Instead, one can model volatility through a latent variable that is stochastic and unobserved. This idea is employed in Stochastic volatility (SV) models, which are special types of state-space models (Tsay, 2005, Chap. 11.2).

The initial equation is similar as the ARCH Equation 1.3 and GARCH Equation 1.4, where the log of returns p_t is described by the mean equation at time t and shock at time t. For SV models, the log returns are assumed to be centered and

$$p_t = \sigma_t \epsilon_t, \tag{1.5}$$

where $\epsilon_t \sim N(0, \sigma_\epsilon^2)$. The main difference is that the shock is not described deterministically by squared historical shocks and conditional variance, but rather by a latent variable h_t that represents the natural logarithm of variance. This is because variance can be only non-negative, and modeling the log variance enforces this assumption. It can then be described by

$$\ln \sigma^2 = h_t.$$

The volatility can be expressed by

$$\begin{aligned} \ln \sigma^2 &= h_t \\ \sigma^2 &= \exp(h_t) \\ \sigma &= \exp\left(\frac{h_t}{2}\right). \end{aligned}$$

Plugging this into 1.5, the simple SV model becomes clear and

$$p_t = \exp\left(\frac{h_t}{2}\right)\epsilon_t.$$

The latent variable is described by an autoregressive process

$$h_{t+1} = \mu + \phi(h_t - \mu) + \eta_t,$$

where μ represents the average log volatility, ϕ the log volatility persistence and $\eta \sim N(0, \sigma_\eta^2)$. For the latent variable to be stationary, it is sufficient that $|\phi| < 1$ (Kim et al., 1998). The addition of another random shocks in the form of η_t makes the SV model more flexible and this can capture more complex relationships. The estimation of latent variable h_t is non-trivial and several methods have been developed, including Kalman filtering and Markov Chain Monte Carlo methods. Similar to GARCH models, the simple SV model is unable to capture the effect of positive and negative spikes (Tsay, 2005, Chap. 3.12). However, additional extensions have been developed to capture the asymmetric property in financial TSD.

2. Bayesian statistics

To describe bayesian statistics, it is useful to first summarize frequentist statistics. The frequentist approach to data analysis has been taught, developed, studied and used for the majority of the 20th century. Some of the greatest statisticians, such as Karl Pearson or Sir Ronald Alymer Fisher, have been prominent figures and authors that applied the frequentist approach and have spoken against the use of bayesian statistics. Frequentist, sometimes referred to as objectivist statistics is based on the idea that the probability of event A can be expressed by a relative frequency. To obtain the frequency, it is needed to observe event A in a large number of independent trials, in the same environment. The observed relative frequency of event A can be assumed to be an estimate of probability of event A. This idea is supported by statistical laws, such is the Law of large numbers (LLN), that states that with an increasing sample size, the relative frequency of event A converges to the probability of event A (Hebák, 2013, Chap. 6).

Frequentist approach implies that probability is objective and that some true population parameter needs to be estimated in an environment where the conditions for a large number of independent trials are the same. The estimated probability is a property of the event and is objective. If one wishes to estimate the probability of some event, the estimate (process of repeated independent repeated experiments) ignores any prior knowledge or whether the probability can be measured. Loosely speaking, the main interest is in the relative frequency (Hebák, 2013, Chap. 6).

This approach has some ideological and computational flaws. The major assumption of independent and identical trials is often impossible to implement in practice. Both independence and same conditions are often impractical in real data analysis. In general, every studied event is different and large number of replications can mean different number in different fields. Ideally, the number of replication should be infinite, but that is not possible to employ. The idea that the studied probability is completely unknown is also very restrictive and more often than not, a researcher knows *something* about the studied event. In a simple coin toss, it is clear that the estimated parameter is in $\langle 0, 1\rangle$ and in researching the average age of population[1], it is quite clear that the estimate is between 18 and 65. Finally, it is assumed that the probability for every independent and identical trial is the same, which is rarely true. Although there is some sensible criticism for frequentist statistics, it is important to highlight that this does not contradict the LLN or other statistical laws. The assumption for such laws is often the independence and identical distribution, that however often do not hold in practice (Hebák, 2013, Chap. 6).

Finally, another property of statistical research in the frequentist mindset is that it is not possible to create probabilistic statements about the estimates. It would be incorrect to

[1]Of course, this depends on the research question. If the researches is tasked to estimate the average age in a university, the range will be much narrower. In a home for the elderly, the range can be wider and much higher.

interpret an estimated $(1-\alpha)\%$ confidence interval that it contains the true parameter with $(1-\alpha)\%$ probability. The true parameter is some fixed number, and the probability that an estimated confidence interval contains the parameter is either 0 (if it lies outside) or 1 (if it lies within) (Hebák, 2013, Chap. 6). The correct interpretation of a confidence interval is that $(1-\alpha)$ % of created intervals will contain the true population parameter while the other α percent will not (Andrade & Fernández, 2016).

Bayesian statistics takes a different approach and instead of estimating the true population parameter, it creates a degree of belief about possible values, that is usually represented using a distribution. The degree of belief can vary for different researchers, because all of them can have individual apriori beliefs about the event. This means that the probability of an event is considered to be subjective and every researcher can have different degrees of belief. Statements such as 'my probability' and 'their probability' about the same event are valid. This does not reject probability axioms defined in the frequentist approach, statistical laws or rules for probability counting (Hebák, 2013, Chap. 6).

When the final estimate is not a relative frequency that depends on identical and independent replications, but it is a degree of belief that is represented by a probability distribution, probability statements are possible. Because the posterior degree of belief is a probability distribution, probabilistic statements are possible and it is trivial[2] to answer statements such as 'What is the probability that the estimated parameter is larger than zero?' or 'What is the probability that the effect in group A is larger than in group B?'. This is also true not only for point estimates, but for intervals as well. In bayesian statistics, confidence intervals are replaced with credibility intervals. Their interpretation is also probabilistic and research question like 'What is the probability that the true population parameter lies between 0 and 10' or 'What is the smallest interval that contains the true parameter value with probability of 0,5?' are also common (Gelman, 2014, Chap. 1.1).

Other than credibility intervals, bayesian statistic offer at least two additional intervals that are useful in statistical inference. The first one is Percentile Interval (PI) that defines a symmetrical center interval of width α. For example, PI of width 0,8 is a percentile interval denoted as $\langle q_{0.1}, q_{0.9} \rangle$. Another common interval is the Highest Density Interval (HDI) that denotes the narrowest interval that contains the posterior parameter θ with probability α (McElreath, 2020, Chap. 3.2).

One benefit of bayesian statistics is that it can be more precise in small samples where frequentist methods struggle because of insufficient repetitions. Issues in statistical inference, such as p-hacking or dichotomization of significance, do not appear in such probability paradigm. Issues regarding frequentist research are out of the scope of this thesis, but they are explored in several will-reviewed books (Vickers, 2010) or papers ((Gelman & Stern, 2006), (Wasserstein et al., 2019)).

In the past, the main drawback of bayesian statistics has been it's computational complexity.

[2]The computation itself is trivial. Definition of a proper bayesian model, its estimation, convergence validation or interpretation might not always be trivial.

Advanced methods, such as hierarchical models, can become very complex and it is not possible analytically express their degree of belief. As such, researchers were limited to some set of special distributions which have analytical solution, but do not have to describe the real world in the most precise way. Thanks to the recent advancements in numerical methods, simulation techniques and availability of computational resources, bayesian methods are more accessible than ever (Hebák, 2013, Chap. 6). Even though the computational possibilities have greatly improved, very complex models can still take relatively long to estimate and do not have to be viable in cases where quick decision making is necessary.

Bayesian statistics is not perfect, and (Depaoli & Van De Schoot, 2017) notes at least three issues that are connected to bayesian data analysis. In small samples, the final degree of belief is heavily affected by the apriori belief that researchers have, and such, the results can vary greatly. It is crucial to report both the final degree of belief about some unknown parameter and the initial belief. Because of the reliance on computational methods, it is important that researchers put emphasis on the estimation process itself and make sure that the final degree of belief that is reported has successfully converged to the proper solution. Finally, researchers might incorrectly interpret the final estimates and a careful review is required[3]. These issues might be prevented by cautious analysis or checklist, such as (Depaoli & Van De Schoot, 2017) or (Gelman et al., 2020), which highlight the most common missteps in bayesian data analysis.

2.1 Bayesian equation and models

Bayesian statistics defines a degree of belief that might be unique to every researcher. It represents the conditional probability distribution of estimated parameters θ given observed data x. This probability distribution is called the posterior distribution which can be computed using the bayesian equation (Marek, 2012, Page 35) as

$$P(\theta|x) = \frac{P(\theta)P(x|\theta)}{P(x)}. \tag{2.1}$$

First is the apriori degree of belief that is individual and is set before observing any new data. One can create their apriori belief on common sense, past research or the domain knowledge. This initial belief is described with a probability distribution that is called the apriori distribution, denoted as $P(\theta)$ where P denotes either probability or density function and θ the unknown parameters. The second main component is based on the collected data and their likelihood, given the apriori degree of belief. The likelihood, denoted as $P(x|\theta)$, represents the conditional probability of data x given the apriori belief about θ. Given some apriori belief, some data is more likely to occur than other. This combination of apriori degree of belief and observed data allows researchers to update their degree of belief based

[3]this issue is however shared and both frequentist and bayesian approach requires careful interpretation and skepticism.

on how it agrees with observations. The final component is called the predictive probability, denoted as $P(x)$ (Hebák, 2013, Chap. 6).

In research, it is often more suitable to talk about data and hypothesis. Equation 2.1 is often rewritten to

$$P(H|D) = \frac{P(H)P(D|H)}{P(D)},$$

but the interpretation is identical. In practice, Equation 2.1 is composed of probability functions and integrals. A simple beta-binomial model is often used to describe the posterior distribution of a parameter that represents probability. It is also often used as an introductory model to bayesian data analysis, for example (Hebák, 2013, Page 155), (Gelman, 2014, Chap. 2.1), or (Kruschke, 2015, Chap. 6).

In the beta-binomial model, parameter θ represents a relative frequency of some event. It is then natural to model the apriori distribution with a $\text{Beta}(a, b)$ distribution as its range is bound to $\langle 0, 1\rangle$. The likelihood of the data, given the apriori distribution $\text{Beta}(\theta|a, b)$, can be expressed by the binomial distribution $\text{Bin}(x|\theta, n)$. Finally, the predictive probability of the data can be rewritten using the law of total probability (Marek, 2012, Page 54) as

$$P(x) = \int_{-\infty}^{\infty} P(x|\theta)P(\theta)d\theta.$$

Combining this information with Equation 2.1 gives the beta-binomial model

$$P(\theta|x) = \frac{\text{Beta}(\theta|a, b)\text{Bin}(x|\theta, n)}{\int_0^1 \text{Beta}(\theta|a, b)\text{Bin}(x|\theta, n)d\theta} = \text{Beta}(\theta|a + y, b + (n - y)). \tag{2.2}$$

The posterior distribution is analytically solvable and is another beta distribution with parameters $a + y$ and $b + (n - y)$, where y represent the number of successes and $(n - y)$ the number of failures (Hebák, 2013, Chap. 6.5.2).

The beta-binomial model is also popular because the posterior distribution is analytically solvable. That is because the beta and binomial distribution are conjugate. If the likelihood function and the apriori distribution are conjugate, the posterior distribution has analytical solution. In the past, selection of specific functions and distributions that are conjugate has been favorable because of this simplicity. If there are better functions or distribution for the analysis that are not conjugate, the posterior distribution might not have analytical solution and thus the analysis becomes difficult. General definition of conjugate prior distributions is offered in (Gelman, 2014, Chap. 2.4).

In real data analysis, it is often needed to use apriori distribution that is not conjugate. In complex models that require multiple apriori distributions, represent some structure in data

or work with latent variables, conjugacy is often not possible and analytical solution does not exist or it is impractical to derive them. To overcome this, sampling and simulation methods have been developed to sample observations from a posterior distribution with the need to know the explicit analytical solutions. These methods do not work with the bayesian Equation in 2.1, but with a simplified version

$$P(\theta|x) \propto P(\theta)P(x|\theta). \tag{2.3}$$

Because the predictive probability, $P(x)$, is constant and does not dependent on the estimated parameters θ, it servers as a normalization constant that can be excluded. This is represented in Equation 2.3 where $\propto$ means that the posterior distribution is proportional to the combination of apriori distribution $P(\theta)$ and likelihood of the data $P(x|\theta)$. The only sufficient requirement for this assumptions is that the apriori distribution is a proper probabilistic distribution (Hebák, 2013, Chap. 6.8).

The choice of the apriori distribution depends on the knowledge that is known before the analysis. This knowledge can come from past research, specific domain knowledge or the meaning of individual parameters. If, for example, it is known that a parameter is surely strictly positive, distributions such as the chi-squared, gamma, exponential or inverse-gamma can be used. Another approach might be to estimate some transformation of a parameter. (McElreath, 2020, Chap. 4.4) chooses to model linear regression with $\ln \beta_j$, which can have any sign. After estimation, the samples from logarithmic posterior distribution can be exponentiated, which yields a positive effect of x_j through $\exp \beta_j$. If the prior information is limited, different approaches exist. One of them is the Jeffreys' invariance principle that states that 'any rule for determining the prior density for $P(\theta)$ should yield an equivalent result if applied to a transformed parameter' (Gelman, 2014, Chap. 2.8). Even though it is popular, it's use in multivariate cases is controversial.

2.2 Markov Chain Monte Carlo methods

To describe posterior distribution that is analytically unknown, bayesian statistics often use a method called Markov Chain Monte Carlo (MCMC) sampling. (Gelman, 2014, Chap. 11) describes MCMC as *'a general method based on drawing values of θ from approximate distributions and then correcting those draws to better approximate the target posterior distribution'*. In statistics, Monte Carlo simulations are used to sample data from a known population distribution, given some input parameters. This allows researchers to compare different methods in the same environment and get insight on what methods might be better in what conditions. Monte Carlo simulations can also be used to estimate the bias, overconfidence, or statistical errors that will be expected when using such methods on samples from the real world. It can be used to measure the precision of point estimates or the expected width of confidence intervals (Hopkins et al., 2024). Monte Carlo chains are simply *'assessing the properties of*

a target distribution by generating representative random values.' (Kruschke, 2015, Chap. 7.4.5). The analytical equation for a posterior distribution is often unknown, and the use of a simple Monte Carlo sampling technique is not sufficient to examine its properties.

Markov Chain is a stochastic process that describes the probability of moving into a different state. The probability of transitioning into a different state at time $t + 1$ is described by a transition matrix T and is dependent only on the current state t. Formally put, the probability that parameter θ transitions to state A is

$$P(\theta_{[t+1]} \in A|\theta_{[0]}, \theta_{[1]}, \dots, \theta_{[t-1]}, \theta_{[t]}) = P(\theta_{[t+1]} \in A|\theta_{[t]}).$$

Such process has important properties that make it suitable in posterior sampling. Chains are able to explore the whole space of possible values and can move between low-density and high-density regions. If the chain is stationary[4], it will converge to a stationary distribution independently on the initial state. Translating these properties into bayesian sampling, properly defined stationary Markov Chains are able to explore the whole posterior distribution without analytical expression (Hebák, 2013, Chap. 6.8). MCMC is combination of Markov Chain samples that are used to describe the true posterior distribution. On these samples, it is possible to compute properties of the posterior distribution, such as the expected value, median, mode, or any interval metric.

There are several methods that use the MCMC methodology. The most popular methods are the Metropolis–Hastings algorithm, Metropolis Algorithm, and the Gibbs Sampler, which are later explored in this thesis. Although MCMC sampling is guaranteed to converge to the posterior distribution, full convergence is the limit of an infinitely long chain. Chains with finite length can describe the posterior distribution with arbitrary precision. It is hence important to not only correctly define the MCMC simulation, but to monitor the quality of convergence. (Kruschke, 2015, Chap. 7.5) defines three main goals when samples are generated using MCMC methods:

1) representativeness,
2) accuracy and stability, and
3) efficiency.

Representativeness of a Markov Chain implies that the posterior distribution is explored fully and that the resulting chains are not affected by the initial starting points. Accuracy of a chain is crucial in determining point estimates, interval estimates or when tails of the distribution are explored. Running the chain multiple times should also give similar[5] results that should be stable. Estimating the posterior should also not take a very long time[6] and the samplers should be efficient.

[4]Stacionarity in Markov Chains is something different than stacionarity in Time series.

[5]The results will never be exactly the same, because the chains are always affected by finite sample size.

[6]This is dependent on the task at hand, and a 'very long time' means something different in a bayesian linear regression, bayesian hierarchical model, estimation of state-space models or in casual DAG simulation.

The quality of chains is very important and is explored in several books, such as (Kruschke, 2015) or (Gelman, 2014), or papers that aim to optimize the bayesian workflow (Gelman et al., 2020). A simple idea is that after the chain has stabilized, the estimated parameters should not change very much. Plotting some central tendency of the estimated parameter against the iteration can be a simple visual check whether the estimate has stabilized. This can be further improved by generating multiple chains and observing whether all chains converge to the same value. Similarly, plotting histograms of the estimated parameters from multiple chains should show high overlap. The difference between chains can also measured with a shrink factor, which represents variance between individual chains. The ideal shrink factor should be very close to 1. Some authors suggest that value greater than 1.1 implies instability and the simulation should be reviewed (Kruschke, 2015, Chap. 7.5.1). These steps validate stability and accuracy, but they can all fail if the resulting stable chains are inaccurate.

To measure the accuracy of created chain, it is appropriate to look at the autocorrelation of generated samples. If the autocorrelation is high, new generated samples are not independent of historical states and the chain can get 'stuck' in a particular place. This means that to explore the whole parameter space and get accurate estimates, the chain needs to be very long. This is often issue for complex models which have high-dimensional parameter space. If a chain exhibits high autocorrelation, there are at least two ways how to correct this. The first is to discard every k-th sample, which reduces the number of times the chain explores a specific region. While this can help with accuracy, the computer still needs to generate all samples before some of them are removed. This reduces the efficiency of sampler, time to convergence and requires larger number of computational power. That's why it is often better to either redefine the model in a more efficient way, or use a sampler/tool that is built for such complexity.

The accuracy of chain can also be measured using Effective Sample Size (ESS) that measures the true sample size that the chain represents. The closer the ESS is to the number of samples that have been generated, the better accuracy of chains there is. It is often computed using the autocorrelation function so the results of autocorrelation analysis and interpretation of ESS should give a similar conclusion. The quality of chains is often heavily influenced by the first generated samples, which can be imprecise and very far from the stationary distribution. That is why the first couple of samples is discarded in a burn-in period. Removing them increases the stability of chain and their accuracy.

2.2.1 Metropolis algorithm

The Metropolis Algorithm (MA) published in 1953 (Metropolis et al., 1953) has been very influential in the popularization of bayesian statistics because of it's simplicity and relative efficiency in bayesian models and is explored in (Gelman, 2014, Chap. 11.2) or (Kruschke, 2015, Chap. 7). The algorithm first needs an initial sample that will be used. Since the stationary distribution of Markov Chain does not depend on the starting value, it can be any possible value whose posterior density is greater than 0. Then, a symmetric proposal

distribution is required that generates new samples, conditional on the last observed sample, to create a Markov Chain. A popular choice can be normal distribution that is centered around the last sample with user defined variance, so

$$\theta^* \sim N(\mu = \theta_{t-1}, \sigma^2).$$

Variance is often denoted as a 'step' that will be taken for the next sample. Large variance implies large jumps between individual samples, while small variance can increase autocorrelation and the chain might not converge efficiently. The choice of variance hence depends on the specific model at hand. After a new sample is generated from the proposal distribution, it is compared to the last sample θ_{t-1} using a simple ratio of posterior densities

$$r = \frac{P(\theta^*|y)}{P(\theta^{t-1|y})}, \tag{2.4}$$

where P denotes the density of the posterior distribution for θ^*, defined in Equation 2.3. If the ratio is greater than 1, it means that the density of the proposed sample θ^* is larger than the density of θ_{t-1} and the sample is accepted with probability of 1. If the ratio is smaller, the density of the proposed sample is smaller than the density of the last sample and the probability of accepting the new sample is r. The decision whether the new sample is accepted or not can defined as

$$\text{prob}(\theta_t = \theta^*) = \min(1, r).$$

If the proposal is accepted, a new sample from the proposal distribution, condition on the new sample θ^* is generated. If not, a new sample θ_t is simply the previous sample θ_{t-1} and a new proposal conditioned on θ_{t-1} is generated.

While this procedure is capable of sampling from the posterior distribution, it might not always be the best choice. One of its main problems is the efficiency in complex parameter spaces where the size of the step needs to be carefully tuned. (Mbalawata et al., 2013) notes that the simple MA can be greatly improved with a careful tuning of the step size. If a Gaussian proposal distribution is used, (Gelman et al., 1996) found that under specific settings, the optimal covariance matrix Σ that is defined by a researcher should be multiplied by a coefficient $\lambda = 2.38^2/d$, where d is the dimension of the matrix.

To make the algorithm itself more efficient, special Adaptive Metropolis Algorithms have been developed that are able to either select an optimal symmetrical proposal distribution or update the covariance matrix of such distribution. This approach allows the user to more efficiently explore the parameter space of complex bayesian models. The final chains are more stable, suffer from smaller autocorrelation and offer better ESS. The description of more adaptive algorithms, discussion of other algorithms and comparison can be found in (Liang et al., 2010).

An example of generated samples from the posterior distribution is in Figure 2.1. These three charts show a simulation of the beta-binomial model defined in Equation 2.2. Because in this model, the true posterior distribution is known, the accuracy of the simulation can be easily compared. The line chart represents the true posterior distribution while the histogram the simulated data.

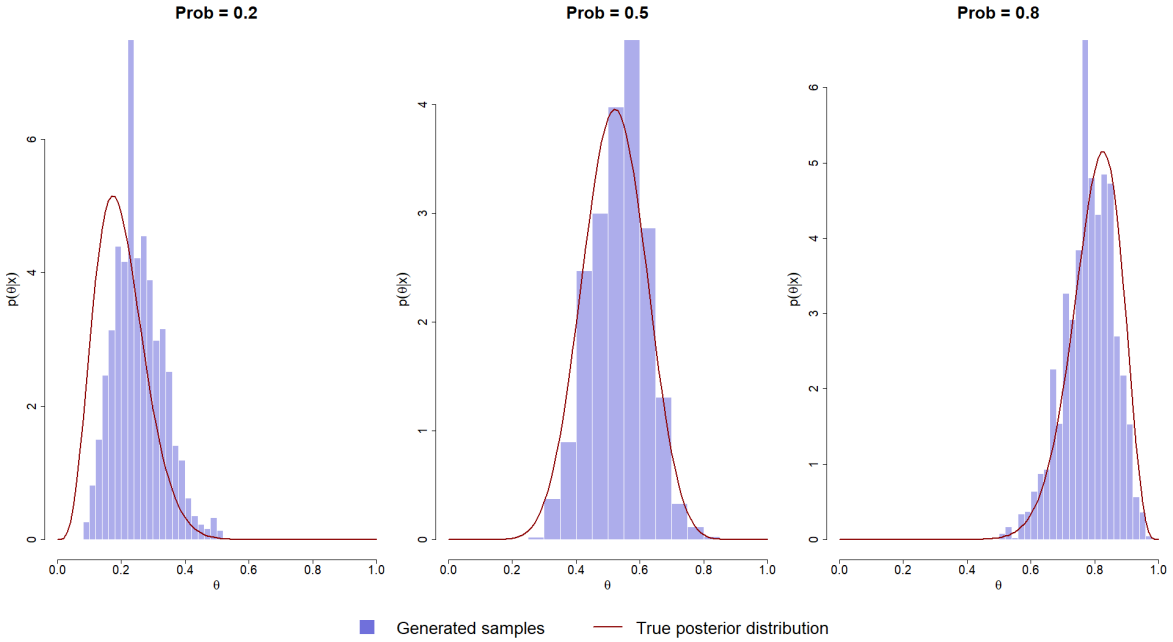

Figure 2.1: Simulation of a beta-binomial model using the Metropolis algorithm.

The simulation generates 10 000 iterations with a burn-in period of 2 000 initial samples. The proposal distribution is symmetric normal distribution $N(\theta_{t-1}, \sigma = 1)$. For this simulation, the generated proposals need to be in $\langle 0, 1 \rangle$, otherwise R throws an error because it is unable to calculate the density of binomial distribution for values greater than 1 or smaller than 0. In smaller and higher true parameter values, the samples do not converge to the true posterior distribution, yet. This may suggest that the variance is large, the symmetric proposal distribution does not fit the model well or that another sampler is needed. The exact code can be found in appendix A.1.

2.2.2 Metropolis-Hastings algorithm

The MA requires the proposal distribution to be symmetric. The Metropolis-Hastings algorithm (MHA) relaxes this assumption and let's the proposal distribution be any proper probability distribution. (Gelman, 2014, Chap. 11.2) denotes that asymmetric proposal distributions can help with convergence and speed in complicated models. Asymmetric proposal distribution is also more appropriate for parameters that itself have asymmetric distribution, such as the variance or standard deviation, whose values are bound at zero (McElreath, 2020, p. 9.2.1).

The main difference is how the ratio of densities is computed. Because of the additional asymmetry, the densities need to be normalized and

$$r = \frac{P(\theta^*|y)/J_t(\theta^*|\theta_{t-1})}{P(\theta_{t-1}|y)/J_t(\theta_{t-1}|\theta^*)} = \frac{P(\theta^*|y)}{P(\theta_{t-1}|y)} \frac{J_t(\theta_{t-1}|\theta^*)}{J_t(\theta^*|\theta_{t-1})}. \tag{2.5}$$

The function J_t in Equation 2.5 denotes the proposal distribution at iteration t. In a simple MA, the ratio of the two conditional proposal distributions is equal to 1 and can be then excluded. In the case of asymmetrical proposal distributions, this is not true, and the ratio needs to be corrected by this factor. The rest of the procedure is the same as for the MA.

In general, (Hebák, 2013, Chap. 6.8) suggest that the only condition for the MHA to generate stationary chains is that the transition probabilities between two states are proportional to the posterior densities, formally put as

$$P(\theta^*|y)J_t(\theta^*|\theta_{t-1}) = P(\theta_{t-1}|y)J_t(\theta_{t-1}|\theta^*).$$

Example of data generated using MHA can be seen in Figure 2.2. The process generates 10 000 iterations with a 2 000 burn-in period. The proposal distribution is in this case the Log-normal distribution with natural logarithm of mean equal to previously accepted sample. The full code can be see in Appendix A.2.

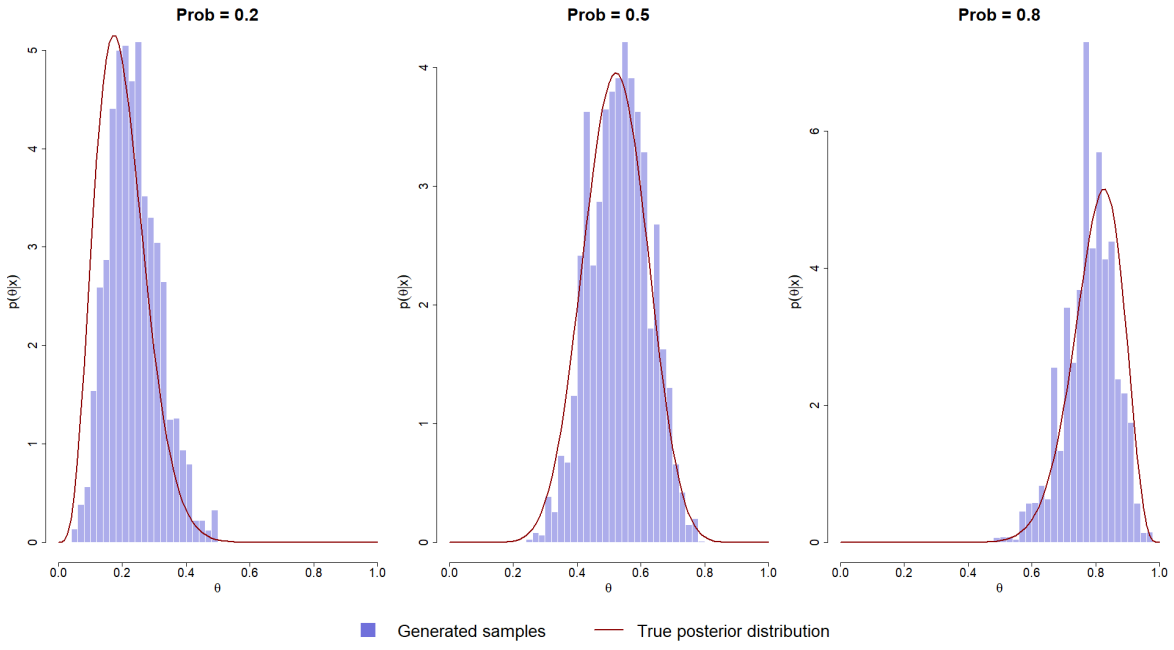

Figure 2.2: Simulation of a beta-binomial model using the Metropolis-Hastings algorithm.

The simulation looks better and it looks like the samples generated using MHA are converging faster than using MA in Figure 2.1.

2.2.3 Gibbs sampling

One of the drawbacks of MA and MHA is the need to fine-tune the proposal distribution so that the samples can efficiently explore the posterior distribution. If the proposal distribution rarely generates values which are common in the posterior, the chains might not converge quickly (Kruschke, 2015, Chap. 7.4.4).

The Gibbs sampler (GS) is a special case of MHA that cleverly sets the proposal distribution, which makes the chain converge more quickly to the posterior distribution. The main idea arises from adaptive MHA that try to update the proposal distribution slightly after some iterations, which might increase the efficiency. Instead of updating the proposal distribution or adjusting it's step size, GS chooses proposal distributions for each parameter individually. The choice is done based on a known, conjugate posterior distribution for each parameter (McElreath, 2020, p. 9.2.1). GS is often better in multidimensional models, where the full posterior distribution is unknown, but the conditional posteriors of parameters is known and values can be sampled from them (Gelman, 2014, Chap. 11.1).

GS is especially useful in hierarchical models, where the conditional probabilities naturally arise. (Kruschke, 2015, Chap. 7.4.4) notes that GS is a special case of MHA. He compares both algorithms to a random walk through the parameter space, where the next step depends only on the current step. While MHA tries to create a step in every direction at once, GS does one step in every direction sequentially. In a multidimensional parameter space $\Theta = [\theta_1, \theta_2]$, MHA generates a new proposal from a proposal multidimensional distribution, so

$$\theta^* \sim N_2(\theta_{t-1}, \Sigma),$$

where Σ is variance-covariance matrix that determines the shape of the proposal distribution. GS generates new proposals conditionally on already accepted samples from conditional posterior distributions. Assuming known posterior distributions $P(\theta_1|\theta_2, x)$ for θ_1 and $P(\theta_2|\theta_1, x)$ for θ_2, the sampler uses them as proposal distributions. First, a new proposal for θ_1 is generated, so

$$\theta_1^* \sim P(\theta_1|\theta_{2,t-1}, x).$$

After the new sample is either accepted or rejected, a new proposal for θ_2 is generated, conditional on the new θ_1 sample and

$$\theta_2^* \sim P(\theta_2|\theta_{1,t-1}, x).$$

While the proposal for θ_1 is generated conditional on $\theta_{2,t-1}$, the proposal for θ_2 is conditional on the sample at time t, $\theta^{1,t}$. This allows the GS to create sequential steps in the posterior parameter space Θ more efficiently. The proposals may be accepted or rejected with

probability of $\min(1, r)$ specified in Equations 2.4 or 2.5. However, (Gelman, 2014, Chap. 11.3) proves that the ratio r is always equal to 1 and a new sample is always accepted. The acceptance-rejection criterion can be skipped, which further improves the speed of convergence as the ratio of two functions does not need to be computed. (Gelman, 2014, Chap. 11.1) describes this process in a general way for arbitrary k-dimensional parameter space and provides graphical comparison of both methods, which can be seen in Figure 2.3. The solid black dots represent a starting point of 5 independent chains that have been used for this simulation. Axis x represents samples of arbitrary parameter θ_1 and axis y parameter θ_2. Note that the figure is a combination of Figures 11.1 and 11.2 from the original source.

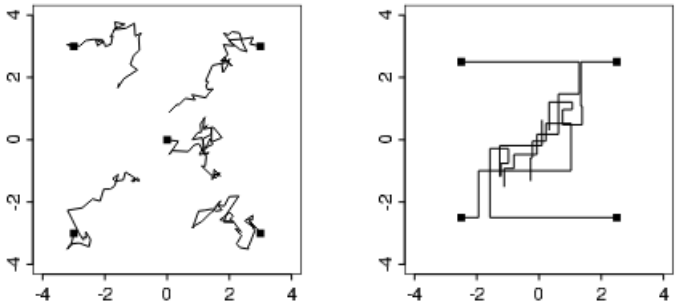

Figure 2.3: Comparison of MHA and GS from (Gelman, 2014)

The left panel shows convergence of MHA, where the step is taken into all directions at once. The right panel shows convergence in steps, where step in either horizontal or vertical direction is taken a at a time.

The efficiency of GS is useful in multidimensional or hierarchical models. (Lu & Chen, 2022) found that a variation of Gibbs sampler is more effective in a multivariate probit model that deals with longitudinal data than other forms of MHA. (Karras et al., 2022) uses GS implemented in Python and PySpark to efficiently sample from the posterior of a general Latent Dirichlet Allocation models. In quantum statistical mechanics, variation of the GS offers smaller variance and autocorrelation compared to the traditional MHA approach (Zhang et al., 2024).

Its main drawbacks, compared to a more flexible MHA is that the conditional distributions need to be known. If the model does not offer conjugate combinations, MHA is more suitable. (Karunarasan et al., 2023) explores this and finds that in some small sample of cases where the conjugate is unkown and computer numerically, MHA is superior. However, they are not able to conclude that MHA is in general superior to GS in abstract multilevel modeling. There are also advanced implementations of MHA that significantly outperform a simple GS. (Mahani & Sharabiani, 2013) develops a multivariate technique that leverages MHA that offers *'6x improvement in sampling efficiency compared to univariate Gibbs'*. The authors argue that this might be because in high dimensional spaces, GS needs to work with complex, high dimensional conditional probability functions that might be hard to compute.

There are also different implementations of GS. (He et al., 2016) define two common ways

how to sample new values. Systematic scan iterates in a predefined cycle of parameters and generates them one-by-one. This means that θ_i is dependent on $\theta j, t$ for $j = 1, \dots, k-1$ and θ_j^{t-1} for $j = k+1, \dots, d$ where d is the dimension of the model. Random scan selects a random parameter at iteration t and samples from a probability distribution conditioned on other latest samples. This is done until all parameters have a new sample for iteration t. (Johnson et al., 2016) describes blocked GS that updates correlated parameters jointly. This can be useful in bayesian linear regression or bayesian generalized linear mixed-effects models.

2.2.4 Hamiltonian Monte Carlo method

MCMC methods leverage a random walk through the parameter space to create samples that are arbitrary close to the posterior distribution. In large, high dimensional models, this approach might struggle because the posterior parameter space. The structure might be so complex that a random walk will take a very long time to explore all possible states, which is not efficient. Even with advanced techniques that implement adaptive steps or change the proposal distribution, the convergence time is not suitable (Gelman, 2014, Chap. 12.4). This complexity might also prevent MCMC methods to adequately sample from regions with low density, and such, tails of the posterior density might burdened with a relatively high MC error (Kruschke, 2015, Chap. 14.1).

Hamiltonian Monte Carlo (HMC) methods do not work with random walk, but with Hamiltonian mechanics. They offer a link between classical mechanics and quantum mechanics. This domain is out of the scope of this thesis, but is explored in some physical textbooks, such as (Lowenstein, 2012). Mathematical explanation of HMC is offered in (Gelman, 2014, Chap. 12.4), while an intuitive introduction to the topic is offered in (Kruschke, 2015, Chap. 14.1) or (McElreath, 2020, Chap. 9.1).

HMC does not implement random walk, but utilizes a set of momentum variables ϕ_i for every estimated parameter θ_i, $i = 1, \dots, d$. Momentum variables represent a 'jump' between every new sample and is completely independent of the estimated parameter θ. Then, the probability distribution of jumps $P(\phi)$, often called the jumping distribution, is combined with the posterior distribution, and

$$P(\phi, \theta|x) = P(\phi)P(\theta|x)$$

is estimated. The jumping distribution is usually set to be a multivariate normal distribution $N_d(\mu, M)$, where M denotes a variance-covariance matrix, often called a mass matrix. If M is a diagonal unit matrix, the distribution is a simple multivariate standardized normal distribution. To make the algorithm more efficient, implementations with different mass matrix or with a different distribution exist.

HMC also works with the gradient of the posterior log-density, which is often computed analytically using mathematical software. The gradient can be defined as

$$\frac{\partial \log p(\theta|x)}{\partial \theta} = \left(\frac{\partial \log p(\theta|x)}{\partial \theta_1}, \dots, \frac{\partial \log p(\theta|x)}{\partial \theta_d} \right).$$

First, the HMC algorithm draws initial values for ϕ from the jump distribution. Then, parameters θ and jumps ϕ are simultaneously updated in multiple iterations using L 'leapfrog steps', where each step is comprised of two parts. First, the jumps are updated with the gradient as

$$\phi_{t+1} = \phi_t + \frac{1}{2}\epsilon \frac{\partial \log p(\theta|x)}{\partial \theta},$$

where ϵ denotes a hyperparameter that scales each step. Then, a new proposal is computed as

$$\theta_{t+1} = \theta_t + \epsilon M^{-1}\phi.$$

This process of leapfrog steps is repeated until the desired number of iterations. After L iterations, new proposed parameter θ^* and steps ϕ^* are obtained. New proposals are used to compute ratio

$$r = \frac{P(\theta^*|x)P(\phi^*)}{P(\theta_{t-1}|x)P(\phi_{t-1})},$$

where the proposals are accepted with probability $\min(1, r)$. After accepting or rejecting a new proposal θ^* and storing new sample θ_t, new jumps ϕ are generated from the jumping distribution and the process is repeated. As well as the MCMC algorithms, HMC is guaranteed to converge to a specific stationary distribution that is the posterior distribution.

There are three ways how HMC can be made more efficient. Firstly, the jumping distribution can be fine-tuned for specific models which can increase the efficiency of samples. It can be though of the jumping distribution as a proposal distribution that can affect how efficient new proposals are. Secondly, scales ϵ can be tuned to allow greater or smaller jumps and steps. Finally, the number of leap frogs L can be optimized. As with MCMC, these parameters can be either set manually or adaptively. As (Gelman, 2014, Chap. 12.4) notes, changing the parameters mid-inference can lead to instability and the adaptive phase should be run only during the warmup phase.

Even though the method offers higher accuracy in small samples, it is computationally more costly, and MCMC methods could be better in relatively simple, non-complex models. It is also limited to continuous parameters and imputation of discrete probability distributions

is not possible[7] (McElreath, 2020, Chap. 9.3). HMC also takes ideas from GS, specifically from the blocked approach to making samples. HMC is able to sample highly correlated parameters jointly in blocks, which can make samples more efficient.

2.2.5 No-U-Turn Sampler

No-U-Turn Samples (NUTS) is a special adaptive technique applied in HMC proposed by (Hoffman & Gelman, 2011). Authors note that the cost of a single independent sample for a d-dimensional model in HMC is $O(d^{5/4})$, which is much higher than the cost of a single MA sample $O(d^2)$. The efficiency of HMC can be greatly affected by two hyperparameters; the scale factor ϵ and the number of leap frogs L. Authors argue that current adaptive algorithms are able to fine-tune parameter ϵ during the warmup phase, however, optimizing L is more difficult. They propose a new NUTS that eliminates the need to manually set the number of steps L, which can increase the widespread use of HMC.

The main idea that authors propose is a *'criterion to tell us when we have simulated the dynamics for long enough, i.e., when running the simulation for more steps would no longer increase the distance between the proposal θ^* and the initial value of θ_{t-1}'.* The criterion that authors propose is

$$(\theta^* - \theta^{t-1})\phi. \tag{2.6}$$

The distance between proposal θ^* and the initial state θ^{t-1} is multiplied by the current momentum ϕ. Ideally, the algorithm should run until criterion in Equation 2.6 is less than zero and the proposal are 'coming back in a circle'[8]. This condition is not computationally feasible and it might not always be possible. Authors propose a new slice variable u, that is used during sampling and solves this issue. Detailed explanation with additional proofs, algorithm description and implementation are outlined in the original paper.

(Hoffman & Gelman, 2011, Chap. 4.4) details that NUTS is at least as good as a vanilla HMC, which makes NUTS often the preferable choice. (Wu et al., 2018) suggest that the NUTS has become the standard sampler for bayesian models.

2.3 Variational inference method

The important characteristics of MCMC methods, such as HA, MHA, HMC or NUTS, is that they converge to the true posterior distribution. If the convergence is sufficiently fast and accurate, generated samples are representative enough that they can be used for statistical

[7] There are implementations that are able to do this. For example, a restructured model in Stan is able to identify such model. Examples are provided in (McElreath, 2020, Chap 15., Chap. 16).

[8] Although not stated in the original paper, this might be the motivation for the name *No-U-Turn*.

inference. The speed of sampling might sometimes be problematic and even with the most advanced samplers, the inference might take a relatively long time. It may happen that the speed is more crucial than the accuracy, and researchers may prefer speed and quick inference with no guarantees that the model represents the exact posterior distribution. This is the case in real-time data analytics or on-demand predictions, where the speed is significantly more important than the accuracy.

The issue can be solved with samplers based on Variational Inference (VI). Variational inference transform the sampling problem from a stochastic processes to an optimization task. Instead of sampling a large number of samples whose stationary distribution is the true posterior, VI tries to find a distribution that is close to the posterior distribution measured by a Kullback–Leibler Divergence (KLD). It is a measure of how two probability distribution are simmilar to each other. In bayesian statistics, it measures the distance between the true posterior $P(\theta|x)$ and arbitrary distribution function q. Interestingly, KLD naturally raises in many statistical methods. One example can be the maximum likelihood estimation that naturally minimizes the KLD (Ranganath, 2017).

(Kullback & Leibler, 1951) define KLD as

$$KL(f_1(x)||f_2(x)) = \int f_1(x) \log\left(\frac{f_1(x)}{f_2(x)}\right) dx \tag{2.7}$$

Applying Equation 2.7 to IV and reversing the fraction in $\log(\cdot)$[9], the equation transforms to

$$\mathrm{KL}(q(\theta)||P(\theta||x)) = -\int q(\theta) \log\left(\frac{P(\theta|x)}{q(\theta)}\right) d\theta. \tag{2.8}$$

The smaller the KLD in Equation 2.8 is, the closer is the proposed distribution q to the true posterior $P(\theta|x)$ becomes. There are many different measures that could be used, however, (Sjölund, 2023) show that with KLD, the predictive probability $P(x)$, that is constant for each model, can be rewritten to

$$\log P(x) = \mathrm{ELBO}(q(\theta)) + \mathrm{KL}(q(\theta)||P(\theta||x)),$$

where ELBO denotes the evidence lower bound[10]

$$\mathrm{ELBO}(q(\theta)) = \mathbb{E}_{q(\theta)}\left[\log(x|\theta) - \log q(x)\right]. \tag{2.9}$$

[9] Because $\log(a/b) = -\log(b/a)$

[10] That is because quantity $P(x)$ is often called the *evidence*. In a similar manner, it could be called the predictive lower bound, but that is not common.

Because the KLD is always non-negative, maximizing ELBO($q(\theta)$) is the same as minimizing KLD. This is useful because minimizing Equation 2.9 is easier than Equation 2.8. (Sjölund, 2023) provides examples that show how the calculations are done in practice.

The probability function $q(x)$ depends on hyperparameters that change how the distribution look and (Gelman, 2014, Chap. 13.7) uses notation $q(x|\phi)$, where ϕ denotes such hyperparameters. The goal of VI is to find function q with hyperparameters ϕ that is closest to the true posterior distribution $P(\theta|x)$. Then, new samples are taken from $q(\theta|\phi)$ and classical inference can be conducted. The ideal lower limit of KLD is zero and thus, the distance can get arbitrary close.

There are various ways how to define distribution q. If the parameters θ are assumed to be independent, then

$$g(\theta|\phi) = \prod_{j=1}^{d} g_j(\theta_j|\phi_j).$$

Detailed process of estimation and the proof that the KLD is decreased with time is provided in (Gelman, 2014, Chap. 13.7).

Even though the function $g(\theta|\phi)$ might not be exact, VI is already being used in research projects. (Gefang et al., 2019) use VI and bayesian shrinkage to develop a new method how to estimate *'Vector Autoregressive models that have hundreds of macroeconomic variables, or more'*. The main motivation for VI in this specific paper is the complexity and computational infeasibility with traditional MCMC methods. (Murakami et al., 2025) use bayesian analysis and VI to identify different crystal phases from X-ray diffraction data and their full profile. The main goal of the paper is to find a bayesian method that utilizes VI to reduce the estimation time *'from a few hours to seconds'*.

2.4 Bayesian ARCH models

ARCH model introduced in Equation 1.3 was a 'quantum jump' for modeling of stochasticity (Geweke et al., 2013, Chap. 5). It is natural to model such model in a bayesian framework. Given the equation shock

$$a_t = \sigma_t \epsilon_t,$$

it is necessary to first set apriori distribution for innovation ϵ_t. It is traditionally assumed to have standard normal distribution, and in cases with fatter tails, the student distribution. This implies that the final model will need to estimate either the variance σ_ϵ^2 or degrees of freedom ν_ϵ.

The conditional volatility σ_t is deterministically composed of historical values and does not have any apriori distribution specified. It is composed of intercept α_0 and coefficients α_i for every squared historical residual that is considered, generally for $i = 1, \dots, m$. The apriori distribution for every coefficient need to respect the non-negativity of volatility and conditions set in Section Section 1.3.1.

(Ari & Papadopoulos, 2016, Chap. 1) sets the apriori distribution for α_0 to be Gamma with parameters r and β. The gamma distribution is strictly positive and positively skewed. It represent the instantinuous volatility at time t. For other coefficients $\alpha_1, \dots, \alpha_m$, they assume joint dirichlet apriori distribution with parameters $\omega_1, \dots, \omega_m$. Dirichlet distribution is a multivariate beta distribution whose coordinates in m-dimensional space always sum to one. This fulfills the assumptions for a stable ARCH mode and the individual coefficients are assumed to be independent. Authors set the distribution of ϵ to be normally distributed and conveniently call such model *Normal-ARCH(m)*. Interestingly, authors did not use MCMC methods but a Lindsey approximation that leverages Taylor series. As the authors themselves note, this method is not suitable for complicated models and an be used in a limited manner.

(Kaufmann & Frühwirth-Schnatter, 2002) analyses switching ARCH models and their use of the Dirichlet distribution for coefficients $\alpha_1, \dots, \alpha_m$ is identical, as well as the use of normal distribution with parameter μ_ϵ and σ_ϵ^2 for ϵ. However, authors chose an Inverse-Gamma distribution for intercepts[11] α_0 that is also bound by zero. Compares to a gamma distribution, it has heavier tails and might be more suitable for models where extreme values are to be expected.

Both papers use custom, non-informative priors. Since this model is relatively non-complex, standard MCMC methods are sufficient. The implementation of ARCH model in Stan can be seen in Appendix #TODO.

2.5 Bayesian GARCH models

Extension of ARCH models are generalized ARCH models introduced in Section Section 1.3.2. It's conditional variance, given in Equation 1.4, can be relatively easily defined using bayesian methods. The assumption about random component ϵ is identical to ARCH model, and usually either normal distribution or the student distribution is selected. The main difference lies in additional β coefficients that represent the effect of lagged conditional variance. Requirements outlined in Section Section 1.3.2 also add that for a stable GARCH model, the sum of paired coefficients at lag i need to be less than one. (Geweke et al., 2013, Chap. 5.1) notes that this condition can be met by discarding draws which do not fulfill this assumption.

If the shock ϵ is modeled using Student's distribution, the prior distribution on degrees of freedom ν must be a proper probability distribution that decreases faster than $1/\nu$. Otherwise, the posterior does not integrate (Geweke et al., 2013, Chap. 5.1). Some author use

[11] The model contains multiple intercepts because it *switches* them based on time.

right side of a truncated Cauchy distribution, which decreases faster. However, student distribution could be worrisome if the estimated degrees of freedom are small. Because of this, some central moments do not exists and another approaches are more suitable. In such cases, a mixture of normal normal distributions might be more suitable (Virbickaite et al., 2015, Chap. 2.2).

The apriori distribution for parameters α and β varies. (Bauwens & Lubrano, 1998) uses flat uniform priors and (Ausín & Galeano, 2007) conducts sensitivity analysis and finds that a change from uniform to beta priors does not cause large differences in the posterior distribution. This might indicate that the choice of priors depend heavily on the task at hand.

The estimation of such model requires MCMC methods, because the posterior distribution is not analytically solvable. (Geweke et al., 2013, Chap. 5.1) notes that the analytical conditional posterior distribution for each parameter is not known, and Gibbs sampler is also not adequate for this model. (Virbickaite et al., 2015) also acknowledges this limitation and (Bauwens & Lubrano, 1998) suggest an alternative called Griddy-Gibbs sampler. The estimation of GARCH model may be complex and fine-tuning of specific MCMC algorithms might be needed for an efficient convergence. (Ardia, 2008) offers an R package `{bayesGARCH}` that offers automatic estimation of GARCH$(1, 1)$ model with shock that has Student distribution. This process is fully automatic and uses MHA with fine-tuned proposal distribution made up of auxiliary ARMA processes. Additionally, a great number of new samplers and methods have been devoted to GARCH model estimation. Overview of some of them can be found in (Virbickaite et al., 2015) or (Geweke et al., 2013, Chap. 5.1). Stan code for a general GARCH model can be found in Appendix #TODO.

2.6 Bayesian SV models

Triantafyllopoulos2021_BayesianInferenceState - 7.3.3

3. Practical application

- Introduction to Stan
- Introduction to various R packages
- Use of ARCH/GARCH/SV models on several financial time series

Závěr

References

AL-Najjar, D. M. (2016-05-06). Modelling and Estimation of Volatility Using ARCH/GARCH Models in Jordan's Stock Market. *Asian Journal of Finance & Accounting*, *8*(1), 152. https://doi.org/10.5296/ajfa.v8i1.9129

Andrade, L., & Fernández, F. (2016-12). Interpretation of Confidence Interval Facing the Conflict. *Universal Journal of Educational Research*, *4*(12), 2687–2700. https://doi.org/10.13189/ujer.2016.041201

Ari, Y., & Papadopoulos, A. (2016-12-10). Bayesian estimation of the parameters of the ARCH model with Normal Innovations using Lindley's approximation. *Economic computation and economic cybernetics studies and research / Academy of Economic Studies*, *3*, 251.

Ashby, D. (2006-11-15). Bayesian statistics in medicine: A 25 year review: BAYESIAN STATISTICS IN MEDICINE. *Statistics in Medicine*, *25*(21), 3589–3631. https://doi.org/10.1002/sim.2672

Ausín, M. C., & Galeano, P. (2007-02). Bayesian estimation of the Gaussian mixture GARCH model. *Computational Statistics & Data Analysis*, *51*(5), 2636–2652. https://doi.org/10.1016/j.csda.2006.01.006

Barber, D. (2012). *Bayesian reasoning and machine learning*. Cambridge University Press.

Bauwens, L., & Lubrano, M. (1998-06-01). Bayesian inference on GARCH models using the Gibbs sampler. *The Econometrics Journal*, *1*(1), C23–C46. https://doi.org/10.1111/1368-423X.11003

Bollerslev, T. (1986-04). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, *31*(3), 307–327. https://doi.org/10.1016/0304-4076(86)90063-1

Campbell, J. Y., Lo, A. W., MacKinlay, A. C., & Whitelaw, R. F. (1998-12). THE ECONOMETRICS OF FINANCIAL MARKETS. *Macroeconomic Dynamics*, *2*(4), 559–562. https://doi.org/10.1017/S1365100598009092

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan : A Probabilistic Programming Language. *Journal of Statistical Software*, *76*(1). https://doi.org/10.18637/jss.v076.i01

Cont, R. (2007). Volatility Clustering in Financial Markets: Empirical Facts and Agent-Based Models. In G. Teyssière & A. P. Kirman (Eds.), *Long Memory in Economics* (pp. 289–309). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-34625-8_10

Depaoli, S., & Van De Schoot, R. (2017-06). Improving transparency and replication in Bayesian statistics: The WAMBS-Checklist. *Psychological Methods*, *22*(2), 240–261. https://doi.org/10.1037/met0000065

Erb, C. B., Harvey, C. R., & Viskanta, T. E. (1996). Political risk, economic risk and financial risk. *Risk Management*. https://api.semanticscholar.org/CorpusID:153921542

Fernández, J., & Rodríguez, G. (2020). *Modeling the Volatility of Returns on Commodities: An Application and Empirical Comparison of GARCH and SV Models*. Pontificia Universidad Católica del Perú. https://doi.org/10.18800/2079-8474.0484

Gefang, D., Koop, G., & Poon, A. (2019). Variational Bayesian Inference in Large Vector Autoregressions with Hierarchical Shrinkage. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3321510

Gelman, A., Roberts, G. O., & Gilks, W. R. (1996-05-09). Efficient Metropolis Jumping Rules. In J. M. Bernardo, J. O. Berger, A. P. Dawid, & A. F. M. Smith (Eds.), *Bayesian Statistics 5* (pp. 599–608). Oxford University PressOxford. https://doi.org/10.1093/oso/9780198523567.003.0038

Gelman, A. (2014). *Bayesian data analysis* (Third edition). CRC Press.

Gelman, A., & Stern, H. (2006-11). The Difference Between "Significant" and "Not Significant" is not Itself Statistically Significant. *The American Statistician*, *60*(4), 328–331. https://doi.org/10.1198/000313006X152649

Gelman, A., Vehtari, A., Simpson, D., Margossian, C. C., Carpenter, B., Yao, Y., Kennedy, L., Gabry, J., Bürkner, P.-C., & Modrák, M. (2020-11-03). *Bayesian Workflow*. arXiv: 2011.01808 [stat]. Retrieved 2024-09-03, from http://arxiv.org/abs/2011.01808

Geweke, J., Koop, G., & Dijk, H. K. van (Eds.). (2013). *The Oxford handbook of Bayesian econometrics* (1. publ. in paperback). Oxford Univ. Press.

He, B., De Sa, C., Mitliagkas, I., & Ré, C. (2016). *Scan Order in Gibbs Sampling: Models in Which it Matters and Bounds on How Much* (1). https://doi.org/10.48550/ARXIV.1606.03432

Hebák, P. (2013). *Statistické myšlení a nástroje analýzy dat* (Vyd. 1). Informatorium.
OCLC: 883371397.

Hindls, R., Arltová, M., Hronová, S., Malá, I., Marek, L., Pecáková, I., & Řezanková, H. (2018). *Statistika v ekonomii*.
OCLC: 1066057734.

Hoffman, M. D., & Gelman, A. (2011). *The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo* (1). https://doi.org/10.48550/ARXIV.1111.4246

Hopkins, V., Kagalwala, A., Philips, A. Q., Pickup, M., & Whitten, G. D. (2024-01-01). How Do We Know What We Know? Learning from Monte Carlo Simulations. *The Journal of Politics*, *86*(1), 36–53. https://doi.org/10.1086/726934

Johnson, N. A., Kuehnel, F. O., & Amini, A. N. (2016). *A Scalable Blocked Gibbs Sampling Algorithm For Gaussian And Poisson Regression Models* (1). https://doi.org/10.48550/ARXIV.1602.00047

Karras, C., Karras, A., Tsolis, D., Giotopoulos, K. C., & Sioutas, S. (2022-09-23). Distributed Gibbs Sampling and LDA Modelling for Large Scale Big Data Management on PySpark. *2022 7th South-East Europe Design Automation, Computer Engineering, Computer Networks and Social Media Conference (SEEDA-CECNSM)*, 1–8. https://doi.org/10.1109/SEEDA-CECNSM57760.2022.9932990

Karunarasan, D., Sooriyarachchi, R., & Pinto, V. (2023-10-03). A comparison of Bayesian Markov chain Monte Carlo methods in a multilevel scenario. *Communications in Statistics - Simulation and Computation*, *52*(10), 4756–4772. https://doi.org/10.1080/03610918.2021.1967985

Kaufmann, S., & Frühwirth-Schnatter, S. (2002-07). Bayesian analysis of switching ARCH models. *Journal of Time Series Analysis*, *23*(4), 425–458. https://doi.org/10.1111/1467-9892.00271

Kim, S., Shephard, N., & Chib, S. (1998). Stochastic volatility: Likelihood inference and comparison with ARCH models. *The Review of Economic Studies*, *65*(3), 361–393. Retrieved 2025-04-07, from http://www.jstor.org/stable/2566931

Kruschke, J. K. (2015). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan* (Edition 2). Academic Press.

Kullback, S., & Leibler, R. A. (1951-03). On Information and Sufficiency. *The Annals of Mathematical Statistics*, *22*(1), 79–86. https://doi.org/10.1214/aoms/1177729694

Liang, F., Liu, C., & Carroll, R. J. (2010-07-16). *Advanced Markov Chain Monte Carlo Methods: Learning from Past Samples* (1st ed.). Wiley. https://doi.org/10.1002/9780470669723

Liu, J. (2024). Navigating the financial landscape: The power and limitations of the ARIMA model. *Highlights in Science, Engineering and Technology*. https://api.semanticscholar.org/CorpusID:270499508

Lowenstein, J. H. (2012). *Essentials of Hamiltonian dynamics*. Cambridge University Press.

Lu, K., & Chen, F. (2022-12). Bayesian analysis of longitudinal binary responses based on the multivariate probit model: A comparison of five methods. *Statistical Methods in Medical Research*, *31*(12), 2261–2286. https://doi.org/10.1177/09622802221122403

Mahani, A. S., & Sharabiani, M. T. A. (2013). *Metropolis-Hastings Sampling Using Multivariate Gaussian Tangents* (1). https://doi.org/10.48550/ARXIV.1308.0657

Marek, Luboš. (2012). *Pravděpodobnost* (1. vyd). Professional Publishing.
OCLC: 806200448.

Mbalawata, I. S., Särkkä, S., Vihola, M., & Haario, H. (2013). *Adaptive Metropolis Algorithm Using Variational Bayesian Adaptive Kalman Filter* (3). https://doi.org/10.48550/ARXIV.1308.5875

McElreath, R. (2020). *Statistical rethinking: A Bayesian course with examples in R and Stan* (Second edition). CRC Press.
OCLC: on1130764237.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953-06-01). Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics*, *21*(6), 1087–1092. https://doi.org/10.1063/1.1699114

Murakami, R., Nagata, K., Matsushita, Y., & Demura, M. (2025). *Rapid, Comprehensive Search of Crystalline Phases from X-ray Diffraction in Seconds via GPU-Accelerated Bayesian Variational Inference* (1). https://doi.org/10.48550/ARXIV.2501.09308

Quantitative analysis on economic and financial factors behind international students tourism (study destination choice): Evidence of china. (2019). https://api.semanticscholar.org/CorpusID:216094186

R Core Team. (2023). *R: A language and environment for statistical computing*. manual. Vienna, Austria, R Foundation for Statistical Computing. https://www.R-project.org/

Ranganath, R. (2017). *Black box variational inference: Scalable, generic Bayesian computation and its applications* [Doctoral dissertation, Princeton University]. http://arks.princeton.edu/ark:/88435/dsp01pr76f608w

Shumway, R. H., & Stoffer, D. S. (2017). *Time Series Analysis and Its Applications: With R Examples* (4th ed. 2017). Springer. https://doi.org/10.1007/978-3-319-52452-8

Sjölund, J. (2023). *A Tutorial on Parametric Variational Inference* (1). https://doi.org/10.48550/ARXIV.2301.01236

Timpson, C. G. (2008-09). Quantum Bayesianism: A study. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics*, *39*(3), 579–609. https://doi.org/10.1016/j.shpsb.2008.03.006

Tsay, R. S. (2005). *Analysis of financial time series* (2nd ed). Wiley.

Vickers, A. (2010). *What is a P-value anyway? 34 stories to help you actually understand statistics*. Addison-Wesley.
OCLC: ocn319602638.

Virbickaite, A., Ausín, M. C., & Galeano, P. (2015-02). BAYESIAN INFERENCE METHODS FOR UNIVARIATE AND MULTIVARIATE GARCH MODELS: A SURVEY. *Journal of Economic Surveys*, *29*(1), 76–96. https://doi.org/10.1111/joes.12046

Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019-03-29). Moving to a World Beyond " $p < 0.05$". *The American Statistician*, *73*, 1–19. https://doi.org/10.1080/00031305.2019.1583913

Wooldridge, J. M. (2020). *Introductory econometrics: A modern approach* (Seventh edition). Cengage.

Wu, C., Stoehr, J., & Robert, C. P. (2018). *Faster Hamiltonian Monte Carlo by Learning Leapfrog Scale* (2). https://doi.org/10.48550/ARXIV.1810.04449

Zabell, S. (2022-05-11). Fisher, Bayes, and Predictive Inference. *Mathematics*, *10*(10), 1634. https://doi.org/10.3390/math10101634

Zhang, W., Moeed, M. S., Bright, A., Serwatka, T., De Oliveira, E., & Roy, P.-N. (2024). *Path integral Monte Carlo in a discrete variable representation with Gibbs sampling: Dipolar planar rotor chain* (1). https://doi.org/10.48550/ARXIV.2410.13633

Použité balíčky

Ardia, D. (2008-06-03). *bayesGARCH: Bayesian Estimation of the GARCH(1,1) Model with Student-t Innovations*. Comprehensive R Archive Network. https://doi.org/10.32614/CRAN.package.bayesGARCH

Appendices

A. Simulation algorithms in R

A.1 The Metropolis algorithm

```r
# Posterior density
p <- function(theta, data) {
  prior <- dbeta(theta, shape1 = 1, shape2 = 1)
  likelihood <- dbinom(sum(data), size = length(data), prob = theta)

  return(prior * likelihood)
}

# Observed data
observed_data <- rbinom(25, size = 1, prob = 0.8)
# Number of iterations
number_of_iterations <- 10000
# Generated samples
samples <- numeric(number_of_iterations + 1)
samples[1] <- 0.5

for (i in seq_len(number_of_iterations)) {
    proposal <- rnorm(1, mean = samples[i], sd = 1)
    proposal <- max(min(1, proposal), 0)
    ratio <- p(proposal, observed_data) / p(samples[i], observed_data)
    if (ratio >= 1 || runif(1) <= ratio) {
        samples[i + 1] <- proposal
    } else {
        samples[i + 1] <- samples[i]
    }
}
```

Source code A.1: Manual Metropolis algorithm in R

A.2 The Metropolis-Hastings algorithm

```r
# Posterior density
p <- function(theta, data) {
  prior <- dbeta(theta, shape1 = 1, shape2 = 1)
```

```r
  likelihood <- dbinom(sum(data), size = length(data), prob = theta)

  return(prior * likelihood)
}

# Observed data
observed_data <- rbinom(25, size = 1, prob = 0.8)
# Number of iterations
number_of_iterations <- 10000
# Generated samples
samples <- numeric(number_of_iterations + 1)
samples[1] <- 0.5

for (i in seq_len(number_of_iterations)) {
    proposal <- rlnorm(1, meanlog = samples[i], sdlog = 1)
    proposal <- max(min(1, proposal), 0)
    ratio <- (
      p(proposal, observed_data) * dlnorm(samples[i], meanlog = samples[i], sdlog =
          1)
    ) / (
      p(samples[i], observed_data) * dlnorm(proposal, meanlog = samples[i], sdlog =
          1)
    )

    if (ratio >= 1 || runif(1) <= ratio) {
        samples[i + 1] <- proposal
    } else {
        samples[i + 1] <- samples[i]
    }
}
```

Source code A.2: Manual Metropolis-Hastings algorithm in R

B. Stan models